ge rates of change in longitudinal

ient for repeated measurements in

ston, P. (1990). Analysis of serial
300, 230–235.
York: John Wiley & Sons.
orth Carolina: SAS Institute.
ample size determination. *Journal*

quency of measurement and study

dies with the bacon pig and their

presence of right censoring caused
ical Trials **9**, 32–46.
ring: Parametric and nonparamet-
iation, Biopharmaceutical Section

accepted June 1997.

BIOMETRICS 54, 331–336
March 1998

Power and Sample Size for Stratified Prospective Studies Using the Score Method for Testing Relative Risk

Jun-mo Nam

Biostatistics Branch, National Cancer Institute,
Executive Plaza North, Room 403, 6130 Executive Boulevard,
MSC 7368, Rockville, Maryland 20892-7368, U.S.A.

SUMMARY

We derive the asymptotic power function of the score test for detecting a common relative risk greater than unity from multiple 2 × 2 tables and formulate methods of sample size determination for use when designing stratified prospective studies. The stratified score test is more efficient than the unstratified test when the latter is unbiased.

1. Introduction

A parameter of major interest in cohort studies (e.g., vaccine trials and animal bioassay experiments) is the relative risk, the ratio of two binomial probabilities. Various approximate methods for interval estimation of the relative risk for a single 2 × 2 contingency table have been presented by, e.g., Noether (1957), Katz et al. (1978), Koopman (1984), Mee (1984), Miettinen and Nurminen (1985), and Bedrick (1987). When more than one 2 × 2 table is involved in a study, several authors, e.g., Gart (1985) and Gart and Nam (1988), have suggested interval estimation of a common relative risk and a test of significance for detecting a relative risk larger than unity. Their method, which possesses many desirable statistical properties, was derived using the general theory of Bartlett (1953). Power and sample size related to the test of a significance has not been thoroughly investigated. They are important in assessing power or in determining sample size requirements for designing a stratified study. In Section 2 of this paper, we present the asymptotic power of the stratified score test and a sample size formula for prospective studies and, in Section 3, we compare the stratified and unstratified tests. Section 4 presents a numerical example using actual data. Section 5 contains concluding remarks.

2. Power and Sample Size of Score Test of $\phi = 1$

Consider J pairs of independent binomial variates, x_{0j} and x_{1j}, with corresponding parameters p_{0j} and p_{1j} with sample sizes n_{0j} and n_{1j} for $j = 1, 2, \ldots, J$. Let $q_{ij} = 1 - p_{ij}$ for $i = 0, 1$ and $j = 1, 2, \ldots, J$. Summation is denoted by dots, e.g., $x_{\cdot j} = x_{0j} + x_{1j}$ and $n_{\cdot j} = n_{0j} + n_{1j}$. Ratios of the two binomial parameters are denoted $\phi_j = p_{1j}/p_{0j}$ for $j = 1, 2, \ldots, J$. Under a common ratio across strata, $\phi_j = \phi$ for $j = 1, 2, \ldots, J$. Gart (1985) provides a statistic for testing $\phi = 1$ as

$$z = \sum_{j=1}^{J} \left\{ (x_{1j} - n_{1j}\hat{p}_j)/\hat{q}_j \right\} \bigg/ \left[\sum_{j=1}^{J} \left\{ n_{1j} n_{0j} \hat{p}_j / (n_{\cdot j} \hat{q}_j) \right\} \right]^{1/2},$$

Email address: jn26r@nih.gov

Key words: Asymptotic power; Relative efficiency; Relative risk; Sample size; Score test; Stratified prospective studies.

where $\hat{p}_j = x_{\cdot j}/n_{\cdot j}$ and $\hat{q}_j = 1 - \hat{p}_j$ for every j. The above statistic, based on the likelihood score, was equivalent to that of Radhakrishna (1965), who extended Cochran's statistic (1954) for a common odds ratio to the common relative risk. The test is known to be locally optimal. Let $u_j = (x_{1j} - n_{1j}\hat{p}_j)/\hat{q}_j$, $v_j = n_{1j}n_{0j}\hat{p}_j/(n_{\cdot j}\hat{q}_j)$, and $w_j = n_{1j}n_{0j}/n_{\cdot j}$ for every j. As n_{ij}'s increase for fixed J, Gart's test statistic is asymptotically normal. The asymptotic power of the one-sided test for $\phi = 1$ against a specific value of $\phi (> 1)$ is expressed as

$$Pr\left\{z \geq z_{(1-\alpha)} \mid H_1\right\} = 1 - \Phi(u), \tag{1}$$

where Φ is the cumulative normal distribution,

$$u = \left[\left\{\sum_j \mathrm{E}(v_j)_0\right\}^{1/2} \cdot z_{(1-\alpha)} - \sum_j \mathrm{E}(u_j)_1\right] \Bigg/ \left\{\sum_j \mathrm{E}(v_j)_1\right\}^{1/2},$$

$\mathrm{E}(v_j)_1 = w_j[n_{\cdot j}\{n_{\cdot j} - (n_{1j}\phi + n_{0j})p_{0j}\}^{-1} - 1]$, $\mathrm{E}(v_j)_0 = w_j p_{0j}/q_{0j}$, and $\mathrm{E}(u_j)_1 = n_{1j}n_{0j}(\phi - 1)p_{0j}/(n_{1j}q_{1j} + n_{0j}q_{0j})$ for $j = 1, 2, \ldots, J$. The form (1) is analogous to Nam (1992, equation (2.1)). The approximate sample size required for a specific power, $1 - \beta$, is found by solving the equation

$$\sum_{j=1}^{J} \mathrm{E}(u_j)_1 = \left\{\sum_{j=1}^{J} \mathrm{E}(v_j)_0\right\}^{1/2} \cdot z_{(1-\alpha)} + \left\{\sum_{j=1}^{J} \mathrm{E}(v_j)_1\right\}^{1/2} \cdot z_{(1-\beta)}$$

from (1). Define design fractions as $t_j = n_{\cdot j}/N$, where $N = \Sigma n_{\cdot j}$ and $s_j = n_{1j}/n_{\cdot j}$, so that $n_{1j} = t_j s_j N$ and $n_{0j} = t_j(1 - s_j)N$ for $j = 1, 2, \ldots, J$. The explicit form of the approximate sample size formula is

$$N = \left\{c_0^{1/2} \cdot z_{(1-\alpha)} + c_1^{1/2} \cdot z_{(1-\beta)}\right\}^2 / \{(\phi - 1) \cdot c_u\}^2, \tag{2}$$

where $c_1 = \Sigma t_j s_j (1 - s_j)[\{q_{0j} - (\phi - 1)s_j p_{0j}\}^{-1} - 1]$, $c_0 = \Sigma t_j s_j (1 - s_j)p_{0j}/q_{oj}$, and $c_u = \Sigma t_j s_j (1 - s_j) p_{0j}/\{q_{0j} - (\phi - 1)s_j p_{0j}\}$. Using the definition by Stuart (1954), the asymptotic efficiency of the score test is

$$\begin{aligned}(A.E.)_s &= \left\{\partial\mathrm{E}\left(\sum_{j=1}^{J} u_j\right)/\partial\phi\right\}_{\phi=1}^2 \Bigg/ \left\{\mathrm{var}\left(\sum_{j=1}^{J} u_j\right)_{\phi=1}\right\} \\ &= \left\{\Sigma t_j s_j (1 - s_j)p_{0j}/q_{0j}\right\} N. \end{aligned}\tag{3}$$

Since $s_j(1 - s_j) \leq 1/4$ for every j, the asymptotic efficiency (3) is maximized when $s_j = 1/2$ for every j. Equal sample size allocation to test and control groups within each stratum is the most efficient, which translates into the best in terms of power or sample size. Calculations of approximate sample sizes for stratified prospective studies under a perfectly balanced design for two strata are summarized in Table 1. The sample size required for a specific power of the test has a strong inverse relation with $\phi - 1$. It is, also, highly sensitive to baseline probabilities.

3. Unstratified Score Test

The score test for $\phi = 1$ based on pooled 2×2 tables can be written as

$$z_p = (x_{1\cdot} - n_{1\cdot}\hat{p})/(n_{1\cdot}n_{0\cdot}\hat{p}\hat{q}/N)^{1/2},$$

where $\hat{p} = x_{\cdot\cdot}/N$ and $\hat{q} = 1 - \hat{p}$. Rewrite this statistic as

$$z_p = \hat{t}/\left\{\mathrm{var}(\hat{t})\right\}_{\phi=1},$$

where $\hat{t} = \hat{p}_1 - \hat{p}_0$ with $\hat{p}_i = x_{i\cdot}/n_{i\cdot}$ for $i = 0, 1$, and $\mathrm{var}(\hat{t})_{\phi=1} = (1/n_{1\cdot} + 1/n_{0\cdot})\hat{p}\hat{q}$. The expectation and variance of $\hat{t}$ are $\mathrm{E}(\hat{t}) = p_1 - p_0$ and $\mathrm{var}(\hat{t}) = (p_1 q_1/n_{1\cdot} + p_0 q_0/n_{0\cdot})$, where $p_1 = (\Sigma t_j s_j p_{0j})\phi/(\Sigma t_j s_j)$ and $p_0 = \{\Sigma t_j(1 - s_j)p_{0j}\}/\{1 - (\Sigma t_j s_j)\}$. The test based on pooled data is biased except under a balanced design. The asymptotic efficiency of the score test based on pooled data is

Table 1

Approximate sample sizes required for 80% *power of the efficient score test for* $\phi =$ 1 *against* $\phi = (1.5, 2.0, 3.0)$ *at* $\alpha = 0.05$ *under the perfectly balanced design when two strata are considered* (N = *total sample size and* n = *sample size of each group*)

Baseline probabilities		N (n)		
p_{01}	p_{02}	$\phi = 1.5$	$\phi = 2.0$	$\phi = 3.0$
0.05	0.1	1258 (315)	324 (81)	84 (21)
	0.2	646 (162)	157 (39)	36 (9)
	0.3	376 (94)	84 (21)	15 (4)
0.10	0.2	543 (136)	133 (33)	31 (8)
	0.3	339 (85)	76 (19)	14 (4)
	0.4	212 (53)	41 (10)	—
0.20	0.3	270 (68)	61 (15)	11 (3)
	0.4	182 (46)	36 (9)	—
0.30	0.4	152 (38)	30 (8)	—

$$(A.E.)_p = \{\partial \mathrm{E}(\hat{t})\,/\partial\phi\}^2_{\phi=1}/\{\mathrm{var}(\hat{t})\}_{\phi=1} = \left(\sum_j t_j p_{0j}\right) N \Bigg/ \left\{4\left(\sum_j t_j q_{0j}\right)\right\} \tag{4}$$

under $s_j = 1/2$ for every j. We can show that $(A.E.)_s \geq (A.E.)_p$ (Appendix), i.e., the asymptotic efficiency of the stratified score test is greater than or equal to that of the unstratified test. Equality holds only if $p_{0j} = p_0$ for every j, i.e., no effect of the stratification. Gart (1992) pointed out that the variance of the efficient stratified estimator of ϕ is smaller than the variance of the pooled estimator when the latter is unbiased. Weinberg (1985) reported a similar finding for the case of two strata. These findings are consistent with the above result of the asymptotic relative efficiency of the score test based on stratified data compared to the test on pooled data. The sample size for prospective studies using the stratified test is smaller than or equal to that using the unstratified test under a balanced design.

4. An Example

Innes et al. (1969) tested the tumorigenicity of Avadex (a fungicide) by continuous oral administration to both males (M) and females (F) of two hybrid strains of mice (X and Y). Frequencies of pulmonary tumors among test mice for categories XM, XF, YM, and YF were 4/16, 2/16, 4/18, and 1/15, and their respective controls were 5/79, 3/87, 10/90, and 3/82. We summarize a statistical analysis in Table 2. Relative risks for development of pulmonary tumors among test animals to controls by strain and sex were 3.95, 3.63, 2.00, and 1.82, respectively. Assuming the homogeneity of relative risks, we obtain an initial value of the MLE of a common relative risk (Tarone, 1981) as $\phi^{(0)} = 2.66$. Corresponding estimated tumor rates among controls by strain and sex are $\hat{p}_{01}^{(0)} = 0.075$, $\hat{p}_{02}^{(0)} = 0.039$, $\hat{p}_{03}^{(0)} = 0.100$, and $\hat{p}_{04}^{(0)} = 0.033$ from, e.g., Gart (1985, equation (2.1)). It requires two iterations to converge to the MLE, $\hat{\phi} = \hat{\phi}^{(2)} = 2.65$. The MLEs of the nuisance parameters are identical to those based on initial estimates to three decimal points. The score statistic for testing homogeneity (Gart, 1985) is $X_3^2 = 0.954$ ($p = 0.81$). With these data, there is no evidence to reject the hypothesis of homogeneity of relative risks across strata. Therefore, we can make inference about the common relative risk ϕ using all information available from combining the four 2×2 tables. The 95% confidence interval for ϕ is (1.35, 5.03) using a method by Gart and Nam (1988). The score method for testing $\phi = 1$ is $z = 2.88$ ($p = 0.002$), and the exact test also yields a very high degree of significance. We may conclude that the fungicide tested is tumorigenic for mice. Note that individual 95% confidence intervals are considerably wider and only one of four intervals does not contain one. The Pearson chi-square test with four degrees of freedom is far less sensitive ($p = 0.049$) than the score test. The test based on pooled data is highly significant. However, we are against the use of pooled data in general because it may lead to a statistical fallacy (see, e.g., Simpson, 1951; Blyth, 1972; Bishop, Fienberg, and Holland, 1975).

Suppose that the probabilities of a pulmonary tumor among control mice by strain and sex are $p_0 = 0.06$, $p_{02} = 0.03$, $p_{03} = 0.11$, and $p_{04} = 0.04$. Assuming homogeneity among relative risks, we obtain the sample size required for 80% power of the score test for $\phi = 1$ against $\phi = 2.65$ as $N = 156$ (i.e., $n_{1j} = n_{0j} = 20$) from (2) under a perfectly balanced design, i.e., $t_j = 1/4$ and $s_j =$

Table 2
Statistical analyses of carcinogenesis bioassay experiments

Strain	Sex	Test mice		Control mice		Relative risk	95% CI
		With tumors[a]	Total	With tumors[a]	Total		
X[b]	M	4	16	5	79	3.95	(1.22, 11.97)
X	F	2	16	3	87	3.63	(0.75, 16.49)
Y[b]	M	4	18	10	90	2.00	(0.70, 5.16)
Y	F	1	15	3	82	1.82	(0.27, 11.48)
Total		11	65	21	338		

Homogeneity test: $\chi^2_3 = 0.954$ ($p = 0.81$)
Score test for $\phi = 1$ against $\phi > 1$: $z = 2.88$ ($p = 0.002$)
Combined analysis under a common relative risk: $\phi = 2.65$
95% C.I. for ϕ: (1.35, 5.03)

[a] Pulmonary tumors.
[b] X, strain X = (C57BL/6XC3H/ANF)F1; Y, strain Y = (C57BL/6XAKR)F1.

1/2 for every j. Similarly, for detecting two- and three-fold increases in relative risk with 80% power, we need $N = 411$ ($n_{1j} = n_{0j} = 51$) and 107 ($n_{1j} = n_{0j} = 13$), respectively. The sample size recommended for a standard carcinogenesis testing protocol (Sontag, Page, and Saffiotti, 1976) is shown to be proper for detecting a two-fold or greater increase with good power.

5. Remarks

The pooled estimator of a common relative risk is not consistent except for a balanced design. When a design is balanced, the unstratified score test for detecting an association is unbiased. It is, however, less efficient than the score test based on stratified data. A sample size calculated for a prospective study using the unstratified test is necessarily greater than that using the stratified score test for the same power.

Sample size determination in stratified case–control studies for detecting a common odds ratio greater than unity using Cochran's test has been investigated by, e.g., Woolson, Bean, and Rojas (1986) and Nam (1992). When no effect of stratification is presented, Cochran's test is less efficient than the unstratified test except for a balanced design. It is different from that of the common relative risk.

From Gart and Nam (1988, equation (7.4)), the skewness of the score statistic for testing unity of a common relative risk is zero when a design is balanced. For an unbalanced design, it is negligible for large n_{ij}'s and fixed J. The sample formula derived under normality of the score statistic is a reasonable approximation in the typical range of alternative values in prospective studies.

Acknowledgements

The author would like to thank two referees and the associate editor for their suggestions for improving an earlier version of the manuscript and Mrs Jennifer Donaldson for careful typing of the manuscript.

Résumé

Nous calculons la fonction puissance asymptotique du score-test pour mettre en évidence un risque relatif supérieur à l'unité à partir de tables 2 × 2 et formulons des méthodes de détermination de tailles d'échantillons dans l'élaboration d'études prospectives stratifiées. Le score-test stratifié est plus efficace que le test non stratifié lorsque ce dernier est non biaisé.

References

Bartlett, M. S. (1953). Approximate confidence intervals. II. More than one unknown parameter. *Biometrika* **40,** 306–317.

Bedrick, E. J. (1987). A family of confidence intervals for the ratio of two binomial proportions. *Biometrics* **43,** 993–998.

Bishop, Y. M. M., Fienberg, S. E., and Holland, P. W. (1975). *Discrete Multivariate Analysis.* Cambridge, Massachusetts: MIT.

say experiments

otal	Relative risk	95% CI
79	3.95	(1.22, 11.97)
87	3.63	(0.75, 16.49)
90	2.00	(0.70, 5.16)
82	1.82	(0.27, 11.48)
38		

/6XAKR)F1.

ncreases in relative risk with 80%
= 13), respectively. The sample size
ontag, Page, and Saffiotti, 1976) is
with good power.

ent except for a balanced design.
ting an association is unbiased. It
data. A sample size calculated for
ager than that using the stratified

o detecting a common odds ratio
y e.g., Woolson, Bean, and Rojas
ted, Cochran's test is less efficient
ifferent from that of the common

e score statistic for testing unity of
unbalanced design, it is negligible
optimality of the score statistic is a
ies in prospective studies.

e editor for their suggestions for
r Donaldson for careful typing of

pour mettre en évidence un risque
es méthodes de détermination de
atifiées. Le score-test stratifié est
aisé.

e than one unknown parameter.

tio of two binomial proportions.

Discrete Multivariate Analysis.

Blyth, C. R. (1972). On Simpson's paradox and the sure-thing principle. *Journal of the American Statistical Association* **67,** 364–366.

Cochran, W. G. (1954). Some methods for strengthening the common χ^2 tests. *Biometrics* **10,** 417–451.

Gart, J. J. (1985). Approximate tests and interval estimation of the common relative risk in the combination of 2 × 2 tables. *Biometrika* **72,** 673–677.

Gart, J. J. (1992). Pooling 2 × 2 tables. Asymptotic moments of estimators. *Journal of the Royal Statistical Society, Series B* **54,** 531–539.

Gart, J. J. and Nam, J.-M. (1988). Approximate interval estimation of the ratio of binomial parameters: A review and corrections for skewness. *Biometrics* **44,** 323–338.

Innes, J. R. M., Ulland, B. M., Valerio, M. G., Petrucelli, L., Fishbein, L., Hart, E. R., and Pallotta, A. J. (1969). Bioassay of pesticides and industrial chemicals for tumorigenicity in mice: A preliminary note. *Journal of the National Cancer Institute* **42,** 1101–1114.

Katz, D., Baptista, J., Azen, S. P., and Pike, M. C. (1978). Obtaining confidence intervals for the risk ratio in cohort studies. *Biometrics* **34,** 469–474.

Koopman, P. A. R. (1984). Confidence limits for the ratio of two binomial proportions. *Biometrics* **40,** 513–517.

Mee, R. W. (1984). Confidence bounds for the difference between two probabilities (letter). *Biometrics* **40,** 1175–1176.

Miettinen, O. and Nurminen, M. (1985). Comparative analysis of two rates. *Statistics in Medicine* **4,** 213–226.

Nam, J.-M. (1992). Sample size determination for case–control studies and a comparison of stratified and unstratified analyses. *Biometrics* **48,** 389–395.

Noether, G. E. (1957). Two confidence intervals for the ratio of two probabilities and some measures of effectiveness. *Journal of the American Statistical Association* **52,** 36–45.

Radhakrishna, S. (1965). Combination of results from several 2 × 2 contingency tables. *Biometrics* **21,** 86–98.

Simpson, E. H. (1951). The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society, Series B* **13,** 238–241.

Sontag, J. A., Page, N. P., and Saffiotti, U. (1976). *Guidelines for Carcinogen Bioassay in Small Rodents.* Publication 76-801, National Institutes of Health, Bethesda, Maryland.

Stuart, A. (1954). Asymptotic relative efficiencies of distribution-free tests of randomness against normal alternatives. *Journal of the American Statistical Association* **49,** 147–157.

Tarone, R. E. (1981). On summary estimators of relative risk. *Journal of Chronic Disease* **34,** 463–468.

Weinberg, C. R. (1985). On pooling across strata when frequency matching has been followed in a cohort study. *Biometrics* **41,** 117–127.

Woolson, R. F., Bean, J. A., and Rojas, P. B. (1986). Sample size for case–control studies using Cochran's statistic. *Biometrics* **42,** 927–932.

Received May 1996; *revised November* 1996; *accepted July* 1997.

Appendix

$(A.E.)_s \geq (A.E.)_p$ *Under Balanced Designs*

The ratio of two asymptotic efficiencies, (3) and (4), is

$$(A.E.)_s/(A.E.)_p = \left\{\sum_j (t_j p_{0j}/q_{0j})\right\} \times \left(\sum_j t_j q_{0j}\right) \Bigg/ \left(\sum_j t_j p_{0j}\right). \tag{A.1}$$

The expansion of the numerator of (A.1) is expressed as

$$N_{u\cdot} = \sum_j t_j^2 p_{0j} + \sum_{j<k}\sum t_j t_k (p_{0j} q_{0k}/q_{0j} + p_{0k} q_{0j}/q_{0k}). \tag{A.2}$$

Since $\Sigma_j t_j = 1$ from the definition in Section 2, the denominator of (A.1) is written as

$$D_{e\cdot} = \left(\sum_j t_j\right)\left(\sum_j t_j p_{0j}\right) = \sum_j t_j^2 p_{0j} + \sum_{j<k}\sum t_j t_k (p_{0j} + p_{0k}). \tag{A.3}$$

From (A.2) and (A.3), the difference between the numerator and denominator is

$$\begin{aligned} N_{u\cdot} - D_{e\cdot} &= \sum_{j<k}\sum t_j t_k \{p_{0j}(q_{0k} - q_{0j})/q_{0j} + p_{0k}(q_{0j} - q_{0k})/q_{0k}\} \\ &= \sum_{j<k}\sum t_j t_k \{p_{0j}(p_{0j} - p_{0k})/q_{0j} + p_{0k}(p_{0k} - p_{0j})/q_{0k}\} \\ &= \sum_{j<k}\sum t_j t_k \{(p_{0j} - p_{0k})(p_{0j}/q_{0j} - p_{0k}/q_{0k})\} \\ &= \sum_{j<k}\sum t_j t_k (p_{0j} - p_{0k})^2/(q_{0j}q_{0k}) \geq 0. \qquad \text{(A.4)} \end{aligned}$$

Therefore, $(A.E.)_s \geq (A.E.)_p$ from (A.1) and (A.4). Note that the equality holds when $p_{0j} = p_{0k}$ for every j and k.